PAT-NO: JP403020866A
DOCUMENT-IDENTIFIER: JP 03020866 A
TITLE: TEXT BASE RETRIEVAL SYSTEM

PUBN-DATE: January 29, 1991

INVENTOR-INFORMATION:

NAME	COUNTRY
SAITO, TAMAKI	
FUKUNAGA, HIRONOBU	

ASSIGNEE-INFORMATION:

NAME	COUNTRY
NIPPON TELEGR & TELEPH CORP	N/A

APPL-NO: JP02035832
APPL-DATE: February 16, 1990

INT-CL (IPC): G06F015/40

ABSTRACT:

PURPOSE: To perform retrieval with high accuracy by extracting one of the word of a questionaire, the synonym of the word, and the word or the synonym with similar coupling relation and the word having the coupling relation as a text coinciding with the content of the questionaire as a retrieval request.

CONSTITUTION: At synonym developing step 5, retrieval structure is reinforced by referring to a synonym dictionary 6 for the word in structure(retrieval structure) which becomes reference to be used in the retrieval generated at structure generating step 4, and selecting the word representing meaning similar to the word. A text retrieval part 7 retrieves a text base 8 setting the retrieval structure generated up to the synonym developing step 5 as a sample, and outputs the word coinciding with the retrieval structure that is the sample as a retrieval result. At this time, morpheme analysis and syntax analysis are performed by using a word dictionary 3. In such a way, it is possible to perform the retrieval with high accuracy from the questionaire by natural language for the data base 8 in which text data of natural language is accumulated as a character code string.

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平3−20866

⑤Int. Cl.[5]	識別記号	庁内整理番号	㊸公開　平成3年(1991)1月29日
G 06 F 15/40	510 M	7313−5B	

審査請求　未請求　請求項の数 1　(全6頁)

㊹発明の名称　テキストベース検索方式

㉑特　願　平2−35832
㉒出　願　平2(1990)2月16日
優先権主張　㉜平1(1989)3月7日㉝日本(JP)㉛特願　平1−54460

㉗発明者　斎藤　珠喜　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
㉗発明者　福永　博信　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
㉑出願人　日本電信電話株式会社　東京都千代田区内幸町1丁目1番6号
㉔代理人　弁理士　磯村　雅俊

明細書

1．発明の名称

テキストベース検索方式

2．特許請求の範囲

(1)見出しの単語とその品詞情報，文法情報等を記憶した単語辞書と，自然言語で書かれた文書を蓄積したテキストベースと，自然言語を用いて文章を入力する入力部と，入力された文章を単語に分割(形態素解析)し，分割した単語の品詞情報，文法情報から入力された文章の文法的構造の解析(構文解析)を行う文解析部と，該文解析部の解析結果に基づいて前記テキストベースを検索する手段とを有するテキストベース検索システムにおいて，前記見出しの単語と同義あるいは類義な意味を有する単語を記憶した類義語辞書と，前記テキストベースの文章を形態素解析，構文解析するテキストベース解析部と，該テキストベース解析部による文章解析結果と前記文解析部による入力文の解析結果とを照合する照合部を設けて，入力文中から，検索時に対象となる一つ以上の単語を選別し，該単語間の格関係を基に検索の標本となるべき構造(検索構造)を生成する構造生成ステップと，該構造生成ステップにおいて作成された検索構造を標本として，前記テキストベース解析部による文章と解析結果と前記文解析部による入力文の解析結果とを前記照合部により照合することにより，前記テキストベース中を検索するテキスト検索ステップとを備えたことを特徴とするテキストベース検索方式．

3．発明の詳細な説明

〔産業上の利用分野〕

本発明は，自然言語の文章データを文字コード列として蓄積したデータベース(以下，これを「テキストベース」という)に対する，自然言語による問合せ文から高精度な検索を可能とするテキストベース検索方式に関する．

〔従来の技術〕

従来のこの種の技術としては，例えば，杉山他による「自然言語理解に基づく情報検索システム

IRIS」(情報処理学会自然言語処理研究会資料NL-58-8,1986)に記載されている如く、データとしての各テキストに対して、その内容に適したキーワード(分野名または言葉)を付与することによって各テキストの内容すなわち特徴を表現し、検索時には、利用者の求めるテキストの内容に関連するキーワード(分野名または言葉等)とその論理的結合関係(AND,OR等)を指定し、その検索条件を満足するテキストを抽出するように構成されているものが知られている。

上記文献において説明されている如き、自然言語による質問文を受付けるインタフェースを有する場合も、質問文を解析することによってユーザの検索要求を対応するキーワードに展開し、それらキーワードの間の論理的結合関係を決めて検索を行う。すなわち、自然言語によるインタフェースを有するか否かにかかわらず、前記テキストベースの検索は、キーワード検索となっていた。

また、検索精度を向上させることを狙ったものとして、細川他による「日本語文構造解析による自動インデクシング方式」(情報処理学会論文誌第21巻3号,1980)に記載されている如く、各キーワードに意味的役割(テキスト中での主体,客体等)を付与する方法も提案されているが、検索時の手掛りとしてキーワードを用いることには変わりはない。

〔発明が解決しようとする課題〕

上記従来技術は、いずれも、テキスト中に含まれているキーワードを手掛りにして検索を行うので、検索の精度、すなわち、ユーザの求めるテキストがどれだけ正しく検索できたか、が高くならないという問題があった。ここで、検索精度の尺度としては、一般に 再現率(ユーザの検索要求に関連するテキストの中で、検索された関連テキストの占める割合)と適合率(検索されたテキストの全体の中で検索された関連テキストの占める割合)が用いられる。

すなわち、テキストの内容にふさわしいキーワードを付与するということは、そのテキストの主題,要旨等を表現するような言葉、あるいは、関連する主要な部分を表わす言葉を、そのテキストを代表する言葉として付与するということであるが、ユーザが検索要求時に思い浮かべるような言い方をすべてキーワードとして付与することは、検索時に不要なテキストを多数出力する結果になり、高い検索精度を確保しながら種々の表現に対応することは難かしい。また、補足的な記述中の情報を検索したい場合についても、補足的な部分にキーワードを付与することは一般的にはないので、キーワード検索によって検索することは不可能である。

本発明は上記事情に鑑みてなされたもので、その目的とするところは、従来の技術における上述の如き問題を解消し、キーワード検索に代る、高い検索精度を有し、かつ、補足的に記述されている事柄をも検索可能なテキストベース検索方式を提供することにある。

〔課題を解決するための手段〕

本発明の上述の目的は、見出しの単語とその品詞情報,文法情報等を記憶した単語辞書と、自然言語で書かれた文書を蓄積したテキストベースと、自然言語を用いて文章を入力する入力部と、入力された文章を単語に分割(形態素解析)し、分割した単語の品詞情報,文法情報から入力された文章の文法的構造の解析(構文解析)を行う文解析部と、該文解析部の解析結果に基づいて前記テキストベースを検索する手段とを有するテキストベース検索システムにおいて、前記見出しの単語と同義あるいは類義な意味を有する単語を記憶した類義語辞書と、前記テキストベースの文章を形態素解析,構文解析するテキストベース解析部と、該テキストベース解析部による文章解析結果と前記文解析部による入力文の解析結果とを照合する照合部を設けて、入力文中から、検索時に対象となる一つ以上の単語を選別し、該単語間の格関係を基に検索の標本となるべき構造(検索構造)を生成する構造生成ステップと、該構造生成ステップにおいて作成された検索構造を標本として、前記テキストベース解析部による文章と解析結果と前記文解析部による入力文の解析結果とを前記照合部により

照合することにより、前記テキストベース中を検索するテキスト検索ステップとを備えたことを特徴とするテキストベース検索方式によって達成される。

〔作用〕

本発明に係るテキストベース検索方式においては、テキストベース検索のための検索要求、例えば、日本語による質問文を解析し、テキストベース中のすべての文章の中から、検索要求の内容に合致するものを抽出することを特徴とするものであり、キーワード検索ではなく、テキストベース中のすべての文章を対象として検索要求に合致するか否かをチェックする点が特徴である。

また、従来のテキストベースの検索方法が、キーワード検索に頼らざるを得なかった理由としては、検索時にテキストの意味内容を解析することは、意味の解析自体が非常に困難であること、および、それを実用的な応答時間の中で実現することは不可能であること等が挙げられる。これに対して、本発明に係るテキストベース検索方式においては、テキストからの意味の抽出は行わず、検索要求としての質問文の内容に合致するテキストとして質問文の語およびその類義語とその結合関係(格関係)と同様の、語または前記類義語のうちの一つおよびその結合関係を有するものを抽出することで、処理の高速化を図り、実用的な応答速度を達成するものである。

〔実施例〕

以下、本発明の実施例を図面に基づいて詳細に説明する。

第1図は、本発明の一実施例を示すテキストベース検索方式の概略フローである。図において、1は入力部、10は解析処理部、3は単語辞書、6は類義語辞書、7はテキスト検索部、8はテキストベースを示している。なお、上記解析処理部10は、後述する文解析ステップ2，構造生成ステップ4，類義語展開ステップ5の各処理ステップを実行する機能を有するものである。

上記単語辞書3には、文解析部2における形態素解析および構文解析に用いる情報が記憶されている。単語辞書3の例は、第2図に示す通りで、その内容は、単語の見出しとその単語の品詞および構文解析に必要な文法情報から成る。第2図の例では、文法情報は、付属語についてその付属語が接続できる語の種類(格助詞の場合は「体言」等)を示してあり、「:」より右には、その付属語が接続する語の格情報が示されている。但し、ここでは、表層的な格情報で示されている。

また、上記類義語辞書6には、類義な意味を表わす単語が納められており、後述する類義語展開ステップ5で参照される。第3図にその一例を示す如く、その内容は、単語見出しと、その単語と類似な意味を持つ単語の集まりから成る。

テキストベース8は、検索対象となるべき文章の集まりであり、何等かの手段により計算機が直接取扱えるような状態、例えば、磁気ディスクや磁気テープ等の中に納められたものである。

入力部1は、テキストを検索するための検索要求(質問)を、自然言語の文章によって入力するものであり、キー操作入力，音声入力，文字のパターン認識等の文字符号化処理を介して自然言語の文章が装置に取込まれる。

文解析ステップ2は、入力部1で入力された文章を解析し、文章の文法的構造を決定する。これには、文章を構成する各単語の識別，分解を行う形態素解析と、それらの単語の結び付き方から、文の構造を決定する構文解析とがある。本ステップ2で行う構文解析は、文章中の各用言に対応した格構造を抽出するものである。

なお、上述の構文解析としては、格文法に対応する格構造(格フレーム)を用意して、その文章の内容を抽出するもの、例えば、Fillmore 等によって行われたものが利用できる。この処理の概要については、例えば、長尾著「言語工学」(昭晃堂，昭和58年刊)の記載が参考になる。

構造生成ステップ4は、前述の文解析ステップ2の結果を受けて検索に用いる単語を取出し、それらの単語相互間の関係から、検索に用いるための標準となるべき構造(以下、「検索構造」という)を生成する。この際、同一内容を表わす複数の自

然言語表現が考えられる場合は、後述する如く、その代表たるべき表現への変換を行う。

類義語展開ステップ5は、上述の構造生成ステップ4で生成された検索構造中の単語について、前記類義語辞書6を参照して、後述する如く、その単語と類似な意味を表わす単語を選択し、検索構造を補強する。

テキスト検索部7は、類義語展開ステップ5までで生成された検索構造を標本として、テキストベース8を検索して、標本である検索構造に合致したものを検索結果として出力する。この際、前述の文解析ステップ2と同様に、単語辞書3を用いて形態素解析と構文解析を行う。

上述の如く構成された本実施例のテキストベース検索方式の動作を、以下、入力部1が入力文

「テキストを検索する」

を、後の処理に送った場合を例として説明する。

文解析ステップ2では、入力文に対し、形態素解析および構文解析を行い、入力文を

「テキスト/名詞」

「を/格助詞」

「検索する/動詞」

に分解し、更に、この入力文の動作は「検索」であり、「検索」の対象は「テキスト」であることを決定する。なお、このとき、実質的に同一内容を表わす複数の表現、例えば、能動態と受動態による表現等に関する構文解析結果は、例えば、第4図に示す如く、各入力文対応に出力される。

構造生成ステップ4では、上述の文解析ステップ2の出力結果から、検索対象となる一つ以上の単語と、それら単語間の関係を示す「検索構造」を生成する。すなわち、ＬＩＳＰ言語型の表現で示すならば、

(検索対象 テキスト))

のようになる。

なお、前述の如く、実質的に同一内容を表わす複数の表現がある場合には、その代表となる構造への変換を行う。すなわち、第4図に示す如く、「テキストを検索する」,「テキストが検索される」,「テキストの検索」の各文章からは、第5図に示す

処理により、ともに、

(検索対象 テキスト))

の構造が生成される。

類義語展開ステップ5では、前記類義語辞書6を参照して、上述の検索構造に含まれる単語を類義語に展開する。例えば、類義語辞書6の中に、「検索」の類義語として「探す」、「テキスト」の類義語として「文書」,「文章」があった場合、上述の検索構造は、

((検索 探す)(対象 (テキスト 文書 文章)))

の如く補強される。

次に、テキスト検索部7は、テキストベース解析ステップ71で、テキストベース8中の各文章の文解析を行い、照合ステップ72で、単語が類義語展開ステップ5から引き渡された検索構造と同様な関係で出現するものを、一致した文書として出力する。従って、上述の例では、「文書を探す」,「文書が検索される」は一致したと判定されるが、「テキストで検索する」は、非一致と判定されることになる。

上記実施例によれば、自然言語の文書から成るテキストベースを検索対象とし、自然言語で検索するテキストを指定し、入力文中の各単語の関係を利用して、入力文中で使用された単語を、その類義語まで展開したもので検索を行うことができるようになり、以下の如き効果が得られる。

(1)テキストベースに対する事前の処理が不要となり、これによる情報の欠落を回避できる。

(2)特に専門知識がなくても利用可能になる。

(3)意味的に近いものを検索できる。

(4)多様な入力文に対応可能になる。

なお、前述の実施例は一例として示したものであり、本発明はこれに限定されるものではないことは、言うまでもないことである。例えば、テキストベース解析ステップ71と文解析ステップ72とは同様の機能を実現するものであり、同一のブロック(モジュール)で共用することも可能である。更に、上記テキストベース解析ステップ71と文解析ステップ72との間に、解析処理部10の構造生成部4と同様の、構造生成ステップを有する如く構成

しても良い。

〔発明の効果〕

以上、詳細に説明した如く、本発明によれば、テキストからの意味の抽出を行うのではなく、検索要求としての質問文の内容に合致するテキストとして質問文の語およびその類義語とその結合関係(格関係)と同様の、語または前記類義語のうちの一つおよびその結合関係を有するものを抽出することにより、キーワード検索に代る、高い検索精度を有し、かつ、補足的に記述されている事柄をも検索可能なテキストベース検索方式を実現できるという顕著な効果を奏するものである。

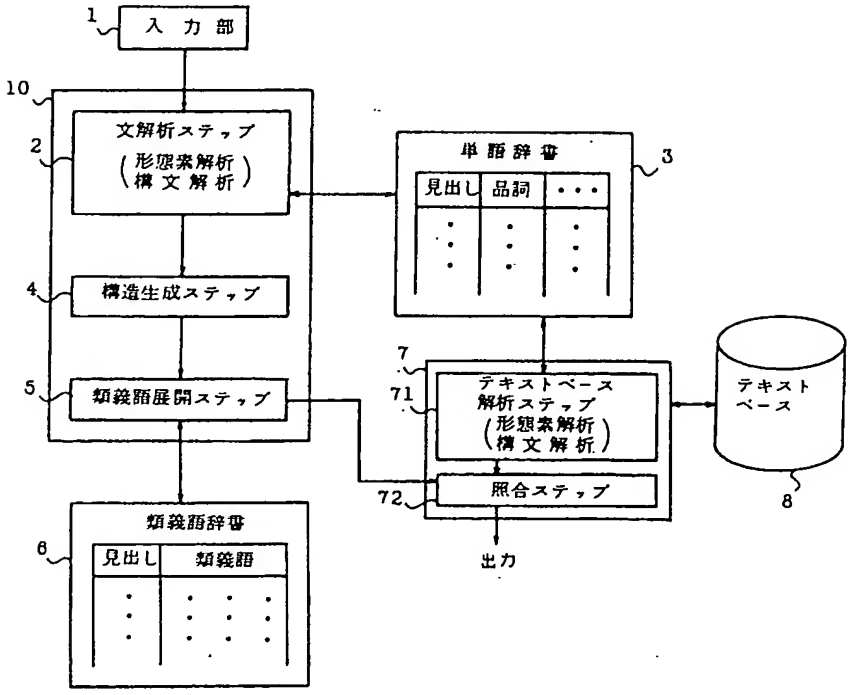
4．図面の簡単な説明

第1図は本発明の一実施例を示すテキストベース検索方式のフローチャート、第2図は単語辞書の内容の一例を示す図、第3図は類義語辞書の内容の一例を示す図、第4図は構文解析結果の一例を示す図、第5図は構造生成ステップの処理の詳細を示すフローチャートである。

1：入力部、10：解析処理部、3：単語辞書、6：類義語辞書、7：テキスト検索部、8：テキストベース、2：文解析ステップ、4：構造生成ステップ、5：類義語展開ステップ、71：テキストベース解析ステップ、72：照合ステップ。

代理人　弁理士　磯村雅俊

第　1　図



第　2　図

（見出し）	（品詞）	（文法情報）
・ ・受動・・・ ・・・ ・・・が・・・ ・・・は・・・ ・・・ ・・・を・・・ ・・・ ・	サ変他動詞 名詞 格助詞 係助詞 格助詞	 体言：主格 体言：主題（主格） 体言：目的格

第　4　図

（構文解析結果）

（検索（対象 テキスト を）（時制 現在）（態能動））
（検索（主体 テキスト が）（時制 現在）（態受動））
（検索（・ テキスト の）（時制 サ変の体言止め））
・
・
・

（・は不確定の意味を表わす）

第　3　図

（見出し）	（類義語）
・ ・ ・ テキスト ・ ・ ・ 検索	 文書，文章 探す

（入力文）

テキストを検索する
テキストが検索される
テキストの検索
・
・
・

第　5　図